

# Notas de Aulas de Redes Neurais Artificiais e de Reconhecimento de Padrões

Prof. Antônio de Pádua Braga

20 de setembro de 2017



# Sumário

0.1	Introdução . . . . .	3
0.2	Regra de Bayes . . . . .	3
0.3	Caracterização do Problema . . . . .	3
0.3.1	Estimador baseado em independência de atributos . . . . .	5
0.3.2	Distribuição Normal Multivariada . . . . .	6
0.3.3	Regra de Bayes . . . . .	7
0.4	Classificador Binário de Bayes . . . . .	8
0.4.1	Exemplos de classificadores binários . . . . .	8
0.4.2	Mistura de densidades . . . . .	9
0.4.3	Estimador por densidade de <i>kernel</i> . . . . .	12
0.4.4	KDE Multivariado . . . . .	13
0.5	Espaço de Verossimilhanças . . . . .	15
0.6	Classificador por matriz de <i>kernel</i> . . . . .	17
0.7	Regra de Bayes . . . . .	18

## 0.1 Introdução

## 0.2 Regra de Bayes

## 0.3 Caracterização do Problema

Considere o problema geral de classificação binária cujo objetivo é discriminar duas classes  $C_1$  e  $C_2$  a partir de um vetor de características  $\mathbf{x} = [x_1, x_2, \dots, x_n]$  e  $N$  amostras das duas classes, caracterizadas pelo conjunto  $D = \{\mathbf{x}_i, y_i\}_{i=1}^N$  em que  $y_i = -1 \forall \mathbf{x}_i \in C_1$  e  $y_i = +1 \forall \mathbf{x}_i \in C_2$ . As características, ou atributos,  $x_1, x_2, \dots, x_n$  que compõem o vetor  $\mathbf{x}$  devem ser representativas do problema e permitir a discriminação das classes  $C_1$  e  $C_2$ . Espera-se que, para um número suficientemente grande de amostras, a estimativa das funções de densidade de probabilidade (fdp)  $P(\mathbf{x}|C_1)$  e  $P(\mathbf{x}|C_2)$  permitam a discriminação das duas classes, caso os atributos selecionados representem fielmente o problema. Atributos não-representativos resultarão em superposição entre as amostras e, consequentemente, entre as densidades estimadas, resultando em um baixo desempenho do classificador. A Figura 1a apresenta um exemplo de amostras de duas classes representadas por meio de atributos não-representativos e, conforme pode ser observado, há uma grande superposição entre as amostras. A Figura 1b, por sua vez, apresenta amostras realizadas por meio de atributos mais representativos

e, por conseguinte, com menor superposição entre as amostras das duas classes. Atributos representativos resultam em maior capacidade de discriminação espacial entre as classes.

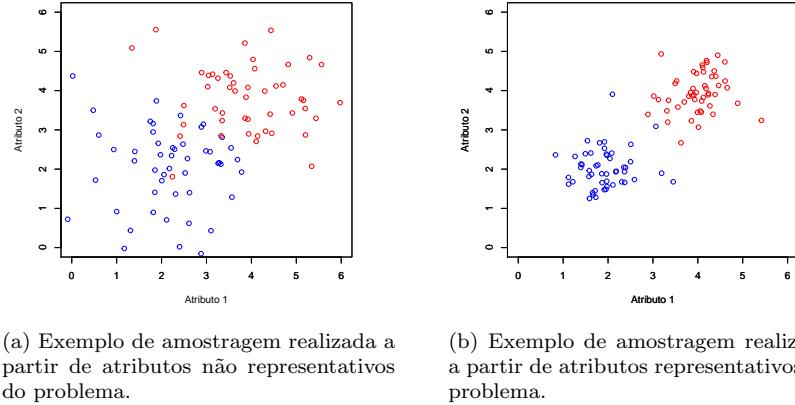


Figura 1: Exemplos de amostragens de dados com atributos representativos e não-representativos.

Caso não haja informação sobre os atributos que compõem  $\mathbf{x}$ , a probabilidade *a priori* de ocorrência de elementos das classes  $C_1$  e  $C_2$  pode ser determinada pelas probabilidades  $P(C_1)$  e  $P(C_2)$ . Considere, por exemplo, uma doença hipotética no Brasil, que ocorre em 20% da população, conforme dados epidemiológicos. Neste caso, a probabilidade *a priori* ( $P(C_2)$ ) de um cidadão brasileiro apresentar esta doença é de 20%, ou seja,  $P(C_2) = 0.2$ , enquanto que a probabilidade de este cidadão não contrair tal doença é de 80% ( $P(C_1) = 0.8$ ). Estas probabilidades *a priori* são estimadas com base no número de ocorrências dos elementos de cada classe, conforme Equações 1 e 2, em que  $N_1$  é o número de elementos de  $C_1$  e  $N_2$  é o número de elementos de  $C_2$ .

$$P(C_1) = \frac{N_1}{N_1 + N_2} \quad (1)$$

$$P(C_2) = \frac{N_2}{N_1 + N_2} \quad (2)$$

A classificação de uma determinada amostra com base somente nas probabilidades *a priori*  $P(C_1)$  e  $P(C_2)$ , no entanto, não considera nenhuma informação sobre a ocorrência de  $\mathbf{x}$  e, portanto, não é capaz de discriminar efetivamente os elementos das duas classes. Com base na informação das distribuições dos atributos é possível estimar as pdfs condicionais, ou verossimilhanças,  $P(\mathbf{x}|C_1)$  e  $P(\mathbf{x}|C_2)$ , e então classificar efetivamente uma determinada amostra com base nos atributos de  $\mathbf{x}$ . A ponderação das verossimilhanças pelas probabilidades *a priori* é um importante elemento da regra de Bayes [LG08], que será descrita nas seções seguintes.

As verossimilhanças  $P(\mathbf{x}|C_1)$  e  $P(\mathbf{x}|C_2)$  podem ser estimadas diretamente a partir das observações de  $\mathbf{x}$ , para cada uma das classes individualmente. Assim,

$P(\mathbf{x}|C_1)$  corresponde à pdf de  $\mathbf{x}$  obtida a partir de todos os elementos da classe  $C_1$ . De maneira análoga,  $P(\mathbf{x}|C_2)$  corresponde à pdf de  $\mathbf{x}$  obtida a partir de todos os elementos da classe  $C_2$ . Conhecidas as duas verossimilhanças e as duas probabilidades *a priori*, um classificador binário, baseado na Regra de Bayes, poderá ser construído.

### 0.3.1 Estimador baseado em independência de atributos

Para o caso particular em que os atributos, ou variáveis, são independentes, ou seja, as correlações entre os atributos de  $\mathbf{x}$  é nula, as verossimilhanças  $P(\mathbf{x}|C_1)$  e  $P(\mathbf{x}|C_2)$  podem ser obtidas diretamente por meio do produto das pdfs marginais, as quais são obtidas individualmente para cada um dos atributos de  $\mathbf{x}$ . Para um problema de duas variáveis, conforme exemplo da Figura 1b, teremos quatro distribuições marginais, sendo duas para cada classe. Assim,  $P(x_1|C_1)$ ,  $P(x_2|C_1)$  são as duas distribuições marginais que compõem a verossimilhança  $P(\mathbf{x}|C_1)$ , a qual pode ser também escrita como  $P((x_1, x_2)|C_1)$ , já que  $\mathbf{x} = [x_1, x_2]^T$ . De maneira análoga,  $P(x_1|C_2)$  e  $P(x_2|C_2)$  são as distribuições marginais componentes da classe  $C_2$ , que compõem a distribuição conjunta de  $x_1$  e  $x_2$  para a classe  $C_2$ , descrita como  $P(\mathbf{x}|C_2)$  ou  $P((x_1, x_2)|C_2)$ . A Figura 2a apresenta as amostras das duas classes, assim como as quatro distribuições marginais representadas sobre cada um dos eixos em cores distintas. Para o caso de independência entre as variáveis  $x_1$  e  $x_2$ , as distribuições conjuntas para cada classe, ou verossimilhanças, podem ser obtidas diretamente por meio dos produtos das marginais, ou seja,  $P((x_1, x_2)|C_1) = P(x_1|C_1)P(x_2|C_1)$  e  $P((x_1, x_2)|C_2) = P(x_1|C_2)P(x_2|C_2)$ . Como pode haver um número grande de variáveis, estas duas distribuições são usualmente representadas tendo o vetor de atributos  $\mathbf{x}$  como argumento, ou seja,  $P(\mathbf{x}|C_1) = P((x_1, x_2)|C_1)$  e  $P(\mathbf{x}|C_2) = P((x_1, x_2)|C_2)$ . Os gráficos das distribuições obtidas por meio dos produtos das marginais da Figura 2a são apresentados na Figura 2b.

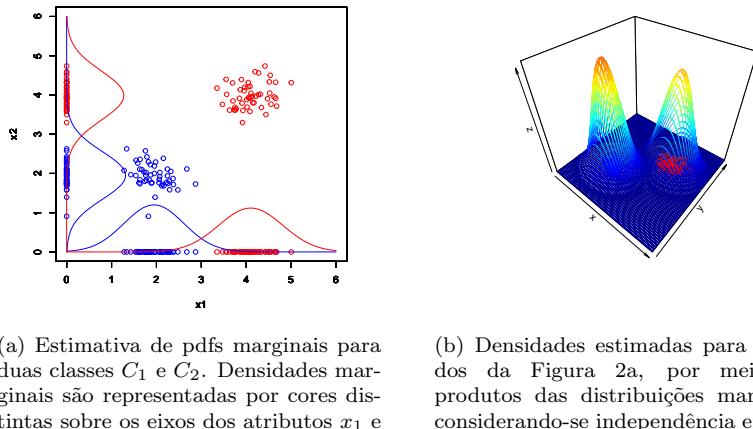


Figura 2: Estimativa de densidades conjuntas considerando-se independência de variáveis.

As expressões matemáticas das pdfs apresentadas na Figura 2b, considerando-se independência entre as variáveis, podem ser obtidas diretamente por meio do produto das expressões das pdfs normais univariadas, para cada um dos atributos e para cada uma das classes individualmente. A expressão da distribuição Normal de uma variável é apresentada na Equação 3, da qual são derivadas as expressões das duas condicionais  $P(\mathbf{x}|C_1)$  e  $P(\mathbf{x}|C_2)$ , apresentadas nas Equações 4 e 5, baseando-se na independência entre os atributos. Como será discutido mais adiante, o classificador de Bayes que considera independência entre as variáveis é chamado de Classificador Ingênuo de Bayes (*Naïve Bayes Classifier*) [HY01].

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad (3)$$

em que  $\mu$  é a média e  $\sigma$  o desvio padrão das amostras.

$$P(\mathbf{x}|C_1) = \frac{1}{2\pi\sigma_{11}\sigma_{12}} e^{-\left(\frac{(x_1-\mu_{11})^2}{2\sigma_{11}^2} + \frac{(x_2-\mu_{21})^2}{2\sigma_{21}^2}\right)} \quad (4)$$

$$P(\mathbf{x}|C_2) = \frac{1}{2\pi\sigma_{12}\sigma_{22}} e^{-\left(\frac{(x_1-\mu_{12})^2}{2\sigma_{12}^2} + \frac{(x_2-\mu_{22})^2}{2\sigma_{22}^2}\right)} \quad (5)$$

em que  $\mu_{11}$  e  $\mu_{12}$  são as médias das pdfs marginais de  $x_1$  para as classes  $C_1$  e  $C_2$ ,  $\mu_{21}$  e  $\mu_{22}$  são as médias das pdfs marginais de  $x_2$  para as classes  $C_1$  e  $C_2$ ,  $\sigma_{11}$  e  $\sigma_{12}$  os desvios-padrão das pdfs marginais de  $x_1$  para as classes  $C_1$  e  $C_2$  e, finalmente,  $\sigma_{21}$  e  $\sigma_{22}$  os desvios-padrão das pdfs marginais de  $x_2$  para as classes  $C_1$  e  $C_2$ .

### 0.3.2 Distribuição Normal Multivariada

Caso os atributos não sejam independentes, a correlação, ou covariância, entre os mesmos deve ser considerada na construção do estimador das pdfs. Para o caso da distribuição Normal de duas variáveis, como aquelas apresentadas na Figura 2b, a expressão geral para a pdf é apresentada na Equação 6. Pode-se observar que, para o caso particular em que a correlação é nula ( $\rho = 0$ ), a Equação 6 se reduz à forma das Equações 4 e 5.

$$p(x_1, x_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} e^{-\frac{1}{2(1-\rho^2)}\left(\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2 - \left(\frac{2\rho(x_1-\mu_1)(x_2-\mu_2)}{\sigma_1\sigma_2}\right) + \left(\frac{x_2-\mu_2}{\sigma_2}\right)^2\right)} \quad (6)$$

em que  $\rho$  é o coeficiente de correlação linear entre as variáveis  $x_1$  e  $x_2$ .

Para o caso geral de múltiplas variáveis, a expressão para a estimativa da densidade conjunta é apresentada na Equação 7.

$$p(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^n |\boldsymbol{\Sigma}|}} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})\right) \quad (7)$$

em que  $n$  é a dimensão de  $\mathbf{x}$ ,  $\boldsymbol{\Sigma}$  é a matriz de covariâncias,  $|\boldsymbol{\Sigma}|$  o seu determinante e  $\boldsymbol{\mu}$  é o vetor de médias das distribuições marginais, conforme apresentado a seguir nas Equações 8 e 9.

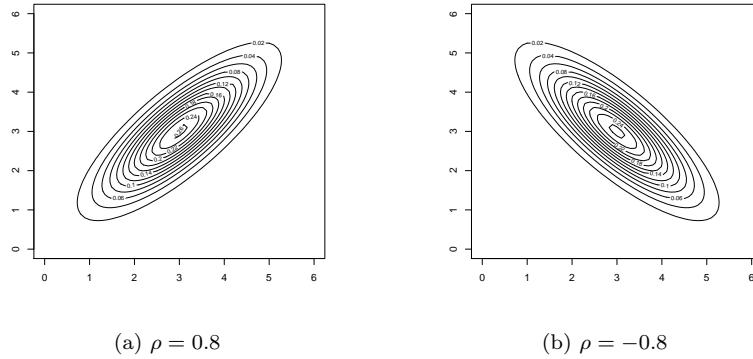


Figura 3: Efeito do coeficiente de correlação na forma da densidade conjunta, observada por meio dos contornos das superfícies.

$$\Sigma = \begin{bmatrix} \sigma_1^2 & \rho_{12}\sigma_1\sigma_2 & \cdots & \rho_{1n}\sigma_1\sigma_n \\ \rho_{21}\sigma_2\sigma_1 & \sigma_2^2 & \cdots & \rho_{2n}\sigma_2\sigma_n \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{n1}\sigma_n\sigma_1 & \rho_{n2}\sigma_n\sigma_2 & \cdots & \sigma_n^2 \end{bmatrix} \quad (8)$$

$$\mu = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_n \end{bmatrix} \quad (9)$$

### 0.3.3 Regra de Bayes

A pdf conjunta  $P(C_i, \mathbf{x})$  entre uma classe arbitrária  $C_i$  e o vetor de atributos  $\mathbf{x}$  pode ser obtida diretamente por meio das probabilidades *a priori* e condicional, conforme Equações 10 e 11.

$$P(C_i, \mathbf{x}) = P(\mathbf{x}|C_i)P(C_i) \quad (10)$$

De maneira análoga, temos:

$$P(C_i, \mathbf{x}) = P(C_i|\mathbf{x})P(\mathbf{x}) \quad (11)$$

A partir das Equações 10 e 11 é possível obter a expressão geral para a Regra de Bayes [LG08], a qual permite estimar a probabilidade *posterior*  $P(C_i|\mathbf{x})$  de uma classe arbitrária  $i$ , considerando-se a ocorrência do vetor de atributos  $\mathbf{x}$ . Assim, igualando-se as Equações 10 e 11, chega-se à expressão geral da Regra de Bayes, a qual é apresentada na Equação 12.

$$P(C_i|\mathbf{x}) = \frac{P(\mathbf{x}|C_i)P(C_i)}{P(\mathbf{x})} \quad (12)$$

Assim, conforme a Equação 12, a probabilidade posterior  $P(C_i|\mathbf{x})$ , dado o vetor  $\mathbf{x}$ , é obtida por meio do produto entre a verossimilhança  $P(\mathbf{x}|C_i)$  e a

probabilidade a priori  $P(C_i)$ , dividido pela evidência  $P(\mathbf{x})$ . A evidência pode ser obtida conforme Equação 13, considerando-se que os eventos que determinam a amostragem de  $\mathbf{x}$  são mutuamente exclusivos, ou seja,  $\mathbf{x}$  será amostrado de somente uma entre todas as classes possíveis. Neste caso, as probabilidades podem ser somadas, conforme Equação 13.

$$P(\mathbf{x}) = \sum_i P(\mathbf{x}|C_i)P(C_i) \quad (13)$$

## 0.4 Classificador Binário de Bayes

A regra de decisão de Bayes, que minimiza o risco de classificação, estabelece que o vetor  $\mathbf{x}$  seja atribuído à classe  $C_j$  de maior probabilidade posterior  $P(C_j|\mathbf{x})$ . Para o nosso problema de duas classes, a regra geral de classificação pode ser simplificada ao atribuir  $\mathbf{x}$  a  $C_1$  se  $\frac{P(C_1|\mathbf{x})}{P(C_2|\mathbf{x})} > 1$  e a  $C_2$  caso contrário. Ao fazer a razão entre as probabilidades posteriores, não há necessidade de realizar o cálculo da evidência  $P(\mathbf{x})$ , pois a mesma é eliminada ao simplificar a expressão resultante. A evidência é apenas um fator normalizador de todas as probabilidades posteriores e, por esta razão, não afeta a ordem das probabilidades posteriores nem a classificação. A regra de Bayes para problemas de classificação binários, obtida por meio da razão das probabilidades posteriores é apresentada na Equação 14.

$$\text{Classe}(\mathbf{x}) = \begin{cases} C_1 & \text{Se } \frac{P(\mathbf{x}|C_1)}{P(\mathbf{x}|C_2)} > k \\ C_2 & \text{Caso contrário} \end{cases} \quad (14)$$

em que  $\text{Classe}(\mathbf{x})$  é a função que atribui  $\mathbf{x}$  a  $C_1$  ou  $C_2$  e  $k = \frac{P(C_2)}{P(C_1)}$ .

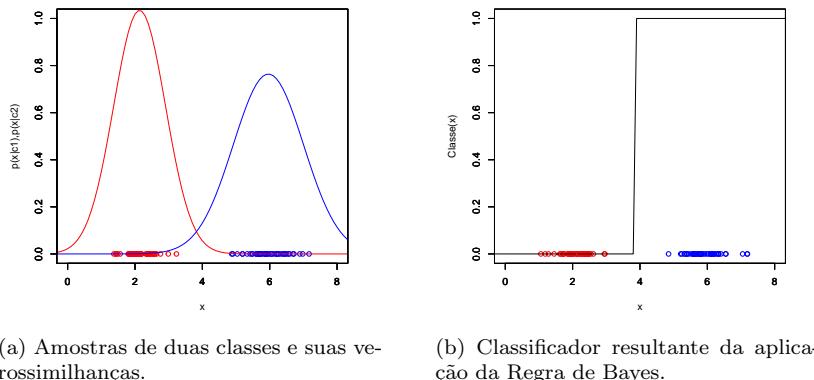
Considerando que a razão  $k$  é uma constante, já que se baseia nos tamanhos das amostras de cada classe, a Equação 14 mostra claramente que a qualidade da classificação depende fundamentalmente da capacidade de discriminação das verossimilhanças. Este conceito reforça a importância da representatividade dos atributos, conforme exemplo das Figuras 1a e 1b.

### 0.4.1 Exemplos de classificadores binários

A aplicação do classificador descrito por meio da Equação 14 requer, portanto, que as verossimilhanças  $P(\mathbf{x}|C_1)$  e  $P(\mathbf{x}|C_2)$  sejam estimadas, assim com as probabilidades a priori  $P(C_1)$  e  $P(C_2)$  de cada classe. Com base nas duas verossimilhanças e nas duas probabilidades a priori, a Regra de Bayes pode ser aplicada de forma que uma amostra arbitrária  $\mathbf{x}$  seja classificada em uma das duas classes  $C_1$  ou  $C_2$ . As probabilidades a priori são obtidas diretamente como a razão entre o número de amostras de cada classe e o número total de amostras, conforme Equações 1 e 2. As verossimilhanças  $P(\mathbf{x}|C_1)$  e  $P(\mathbf{x}|C_2)$  são obtidas a partir das amostras de cada classe individualmente, por meio de modelos de distribuição paramétricos ou não paramétricos.

A Figura 4a mostra um exemplo de um classificador binário univariado em que foram utilizados modelos de distribuições normais para estimar as verossimilhanças. Um teste de normalidade foi inicialmente aplicado aos dados para avaliar se um modelo de distribuição Normal poderia ser aplicado. Ambos os

conjuntos de amostras da classe  $C_1$  quanto aqueles da classe  $C_2$  passaram no teste de normalidade, o que permitiu a utilização de modelos Normais para ambas as distribuições. Foram, então, estimadas as médias e as variâncias das amostras de cada classe, resultando nas distribuições Normais  $P(\mathbf{x}|C_1)$  e  $P(\mathbf{x}|C_2)$  apresentadas na Figura 4a a partir das quais foi possível aplicar a Regra de Bayes e obter a resposta do classificador apresentada na Figura 4b. A resposta em degrau do classificador mostra o seu comportamento na região em que os dados foram amostrados e indica claramente que a aplicação da regra de Bayes resultou em um classificador com limiar rígido próximo à média das médias de cada classe.



(a) Amostras de duas classes e suas ve-  
rossimilhanças.  
(b) Classificador resultante da aplica-  
ção da Regra de Bayes.

Figura 4: Classificador de Bayes de uma variável. Exemplo de dados amostra-  
dos, densidades estimadas para cada classe seguidos da aplicação da Regra de  
Bayes para classificação.

A Figura 5 mostra um exemplo análogo àquele apresentado na Figura 4, porém, agora para a situação em que o problema envolve duas variáveis. Da mesma forma, assumiu-se modelos de distribuição Normais para ambas as classes após as amostras em cada dimensão terem passado em testes de normalidade. Os modelos bidimensionais foram então estimados por meio da Equação 7 e apresentados na Figura 5a. A resposta do classificador resultante é apresentada na Figura 5b.

#### 0.4.2 Mistura de densidades

Nas seções anteriores, exemplos de classificadores Bayesianos foram apresentados considerando-se dados sintéticos amostrados de distribuições Normais, cujas densidades foram estimadas considerando-se normalidade dos mesmos. Em várias situações reais, no entanto, as amostras não obedecem a distribuições normais e, consequentemente, a condição de normalidade não pode ser assumida. Em tais situações, pode-se avaliar a utilização de distribuições alternativas à Normal, adotar modelos não-paramétricos como o KDE (*Kernel Density Estimation*) [DHS01] ou utilizar modelos de misturas de densidades. Nos modelos de mistura, assume-se usualmente um modelo paramétrico, como a distribuição Normal, para cada um dos seus componentes. Assim, a função de densidade é

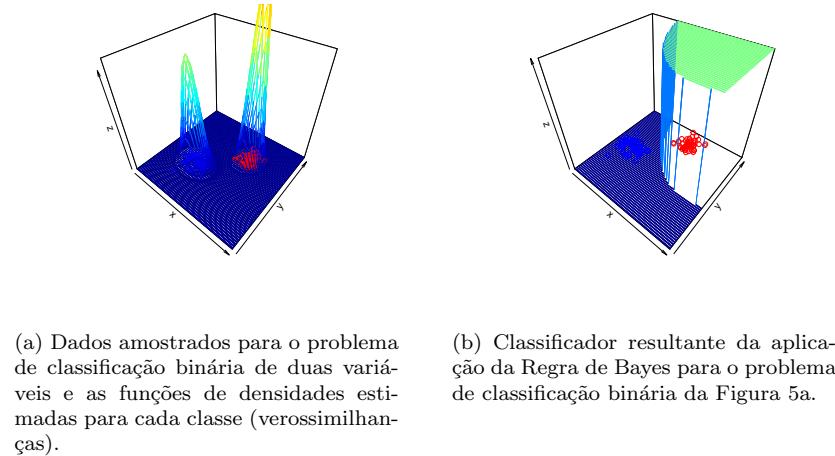


Figura 5: Classificador Bayesiano de duas variáveis. Gráficos apresentam os dados amostrados, as densidades estimadas para cada classe e a resposta do classificador em uma região limitada do espaço de entrada.

estimada por meio de uma combinação linear (mistura) das probabilidades em cada partição, conforme Equação 15.

$$P(\mathbf{x}|S_1, \dots, S_p) = \sum_{k=1}^p \pi_k P(\mathbf{x}|S_k) \quad (15)$$

em que  $\pi_k = \frac{N_k}{N}$  é a probabilidade da partição  $S_k$ ,  $N_k$  é o número de amostras de  $S_k$ ,  $N$  o número total de amostras e  $P(\mathbf{x}|S_k)$  é a probabilidade de  $\mathbf{x}$  considerando-se somente as amostras da partição  $S_k$ .

Como as probabilidades  $\pi_k$  são conhecidas a priori e não dependem dos parâmetros das distribuições, elas correspondem aos coeficientes da combinação linear dos termos  $P(\mathbf{x}|S_k)$ . Assim, para cada partição haverá uma função de densidade associada, de acordo com o modelo assumido pelo projetista. Para o caso de assumir-se distribuições Normais para todas as partições, as probabilidades  $P(\mathbf{x}|S_k)$  podem ser estimadas diretamente por meio da Equação 7, o que resulta na Equação 16.

$$P(\mathbf{x}|S_1, \dots, S_p) = \sum_{k=1}^p \pi_k \frac{1}{\sqrt{(2\pi)^n |\Sigma_k|}} \exp \left( -\frac{1}{2} (\mathbf{x}_k - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1} (\mathbf{x}_k - \boldsymbol{\mu}_k) \right) \quad (16)$$

A divisão do conjunto de amostras nas  $p$  partições das Equações 15 e 16 pode ser feita por meio de métodos de agrupamento (*clustering*) [KR90] *a priori* ou de forma concorrente ao ajuste de parâmetros dos elementos da mistura. Na primeira situação, os dados são inicialmente divididos nas partições  $S_1, S_2, \dots, S_p$  e, posteriormente, um modelo de densidade é estimado para cada uma das partições para que então a mistura possa ser calculada. Métodos de agrupamento, como o K-Médias [Mac67] são usualmente utilizados para realizar a partição dos

dados. Na segunda situação, a divisão do conjunto de amostras nas partições  $S_1, S_2, \dots, S_p$  é realizada de maneira concomitante à obtenção dos parâmetros, ou seja, por meio de um método de Otimização e de uma função-objetivo definida previamente, todos os parâmetros dos modelos são obtidos. Um exemplo de método que é usualmente utilizado dentro desta abordagem é o algoritmo de Maximização da Expectativa (EM - *Expectation Maximization*) [EM77].

A Figura 6 apresenta um exemplo de mistura de distribuições Normais, cujas partições foram obtidas *a priori* por meio do algoritmo K-Médias. Como pode ser observado na Figura 6a, os contornos da função de densidade estimada parecem coerentes com a distribuição das amostras. Para este caso particular, o número de partições adotado coincidiu com o número de modos (quatro) da função geradora. Apesar de o número de partições ter sido obtido diretamente por inspeção visual, o que é possível neste caso particular de duas variáveis, existem métodos quantitativos para avaliar se o número de partições está coerente com um determinado conjunto de amostras [KR90]. A superfície resultante da mistura é apresentada na Figura 6b.

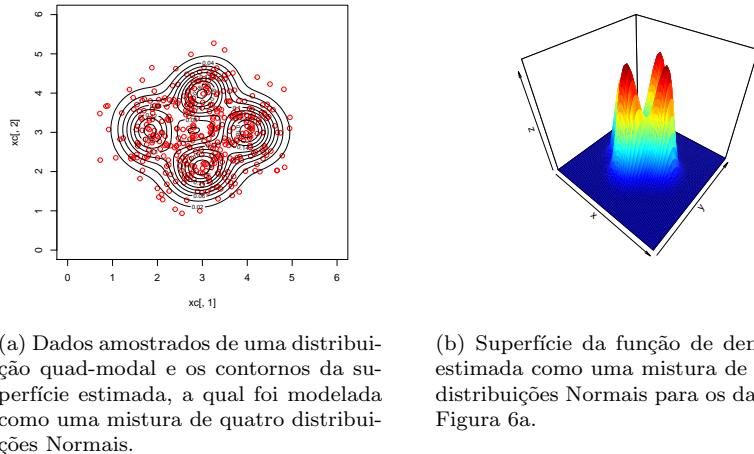


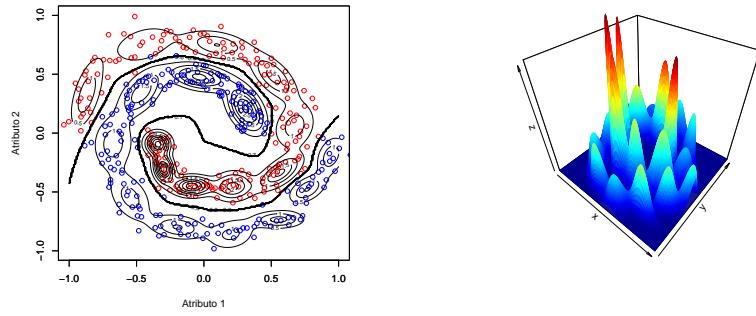
Figura 6: Exemplo de mistura de distribuições Normais estimada conforme Equação 16.

### Classificador bayesiano utilizando mistura de densidades

Os exemplos de classificadores Bayesianos dados até o momento foram unimodais, lineares e foram resolvidos assumindo-se normalidade dos dados. No entanto, boa parte dos problemas reais possui multi-modalidade e tem características de separação não-lineares, como é o caso do problema de *benchmarking* das espirais que será utilizado como exemplo nesta seção. Devido à forma como as amostras das duas classes estão distribuídas, a separação requer um modelo não-linear mais complexo do que aquele resultante da aplicação de funções de densidade unimodais. Para este problema em particular será utilizado um modelo de mistura de distribuições normais, com dez partições para cada classe. A Figura 7a mostra as amostras de cada classe, os contornos de cada mistura,

correspondentes a  $P(\mathbf{x}|C_1)$  e  $P(\mathbf{x}|C_2)$ , e o contorno da superfície de separação resultante da aplicação da Regra de Bayes para o problema. A Figura 7b apresenta as superfícies das misturas, correspondentes às duas verossimilhanças  $P(\mathbf{x}|C_1)$  e  $P(\mathbf{x}|C_2)$ .

Como pode ser observado na Figura 7b a superfície de separação resultante da utilização dos modelos de mistura e da aplicação da Regra de Bayes separa de maneira coerente as amostras das duas classes. Neste caso, o número de partições, ou de elementos da mistura, pode ser estimado com base no desempenho do classificador, caracterizado pela forma da superfície, utilizando métodos de validação [Koh95].



(a) Amostras de dados das duas classes  $C_1$  e  $C_2$ , contornos das misturas e contorno da superfície de separação resultante.

(b) Funções de densidade  $P(\mathbf{x}|C_1)$  e  $P(\mathbf{x}|C_2)$  que resultaram nos contornos da Figura 7a.

Figura 7: Classificador de Bayes por mistura de distribuições Normais para o problema de classificação das duas espirais.

#### 0.4.3 Estimador por densidade de *kernel*

Conforme discutido nas seções anteriores, quando o conjunto de amostras atende aos critérios de normalidade, a distribuição pode ser estimada a partir de parâmetros como média, desvio-padrão e correlação, conforme descrito na Seção 0.3.2. No entanto, nem sempre o conjunto de amostras será bem comportado o suficiente para se enquadrar em modelos paramétricos, como a distribuição Normal ou mesmo misturas de distribuições Normais. Em tais situações adversas é ainda possível estimar uma função de densidade para descrever a função geradora dos dados por meio de modelos não-paramétricos, como o KDE (*Kernel Density Estimation*) [Par62].

A estimativa pelo KDE é realizada através da superposição de funções de densidade centradas em cada um dos pontos do conjunto de amostras, conforme apresentado na Equação 17. No entanto, apesar de a restrição para a função de kernel ser de que ela seja simétrica e com integral unitária, podendo então assumir várias formas, tipicamente utiliza-se a função de densidade Normal, apresentada na Equação 18 como função de kernel.

$$p(x) = \frac{1}{N} \sum_{i=1}^N \frac{1}{h} K\left(\frac{x - x_i}{h}\right) \quad (17)$$

$$K\left(\frac{x - x_i}{h}\right) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x - x_i}{h}\right)^2} \quad (18)$$

onde  $N$  é o número de amostras e  $h$  a abertura das funções normais.

Um exemplo de aproximação de uma densidade Normal univariada e unimodal com o KDE é apresentado na Figura 8a, na qual pode ser observada a proximidade entre as duas estimativas. A Figura 8b mostra um exemplo de densidade estimada com o KDE em que um conjunto reduzido de dados foi gerado a partir de duas distribuições Normais com médias em  $m_1 = 2$  e  $m_2 = 4$ . Para este problema em particular, cada um dos modos da distribuição passaria individualmente no teste de normalidade, não obstante, o conjunto de amostras é caracterizado por uma distribuição bimodal. Uma modelagem paramétrica mais poderia envolver a identificação dos modos da distribuição e modelar cada um deles separadamente, resultando em uma mistura de densidades, conforme discutido na Seção 0.4.2. No entanto, para problemas multidimensionais, esta é uma decisão difícil de ser tomada pelo projetista e uma alternativa para a modelagem neste caso é utilizar o KDE. Assim, como pode ser visto na Figura 8b, com um valor próprio para o parâmetro  $h$  o resultado da estimativa com o KDE é bastante próximo das funções de densidade que geraram os dados. A figura mostra também a estimativa da densidade utilizando-se um modelo de distribuição Normal para todo o conjunto de dados.

Para a modelagem com o KDE é necessário estimar somente o valor de  $h$ , parâmetro único, porém, a sua estimativa envolve a resolução de um dos problemas mais fundamentais em Aprendizado de Máquina, o *equilíbrio entre o viés e a variância do modelo* [GBD92], discutido também em outras seções deste livro.

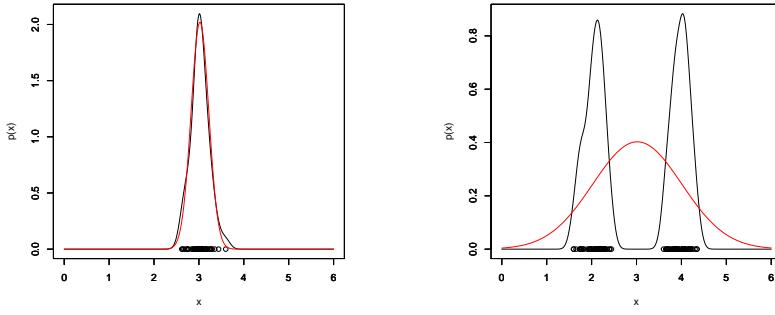
Apesar de haver na literatura vários trabalhos sugerindo regras para estimar o valor de  $h$  para um conjunto de dados, o seu valor é crítico e pode influenciar bastante o resultado final. Para o exemplo da Figura 8b, o valor de  $h$  foi obtido por simples inspeção do gráfico, considerando-se o conhecimento prévio sobre as funções geradoras dos dados. Em situações reais, frequentemente o gráfico não pode ser visualizado devido à dimensionalidade do espaço de entrada e, é claro, devido também à falta de conhecimento prévio sobre as distribuições geradoras, a não ser através das informações estatísticas extraídas dos dados. Assim, em situações reais o valor de  $h$  deve ser estimado diretamente através do conjunto de amostras. De acordo com Silverman [Sil86] uma regra prática para estimar o  $h$  de uma distribuição univariada é apresentada na Equação 19.

$$h \approx 1.06\hat{\sigma}N^{-\frac{1}{5}} \quad (19)$$

em que  $N$  é o número de amostras e  $\hat{\sigma}$  é o desvio-padrão estimado dos dados.

#### 0.4.4 KDE Multivariado

Conforme visto na seção anterior, para um valor apropriado de  $h$  é possível estimar a densidade de uma variável através da Equação 17. Para o caso de o problema envolver mais de uma variável, a maneira mais direta de estimar



(a) Comparação entre as estimativas de uma distribuição Normal realizadas com o KDE, utilizando a Equação 19 para estimar o valor de  $h$ , e com a função de densidade Normal.

(b) Estimativa com o KDE de uma função de densidade bimodal, composta por funções Normais geradoras com médias em  $x = 2$  e  $x = 4$ . A figura apresenta também a estimativa da função de densidade utilizando-se função a função de densidade Normal.

Figura 8: Classificador de Bayes por mistura de distribuições Normais para o problema de classificação das duas espirais.

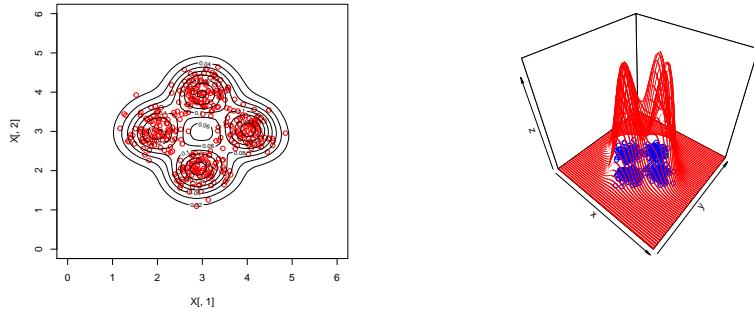
a densidade conjunta é construir um estimador ingênuo, assumindo-se a independência entre as variáveis de entrada, de maneira análoga ao exemplo da Seção 0.3.1. A densidade resultante é obtida então por meio do produto das densidades para cada uma das variáveis independentes. Assim, considerando-se independência entre as variáveis e o mesmo raio  $h$  para todas as dimensões, a estimativa de densidades pelo KDE Gaussiano (Normal) em um determinado ponto arbitrário  $\mathbf{x}_i = [x_{i1}, x_{i2}, \dots, x_{in}]^T$  pode ser obtida por meio da soma dos produtos acumulados em todas as dimensões para todos os padrões do conjunto de amostras, conforme Equações 20 a 23 que se seguem. A Figura 9b mostra uma estimativa de densidade para um problema em que as amostras são caracterizadas por uma distribuição multimodal.

$$p(\mathbf{x}_i) = \frac{1}{N} \sum_{j=1}^N \frac{1}{h} K\left(\frac{x_{i1} - x_{j1}}{h}\right) \frac{1}{h} K\left(\frac{x_{i2} - x_{j2}}{h}\right) \dots \frac{1}{h} K\left(\frac{x_{in} - x_{jn}}{h}\right) \quad (20)$$

$$p(\mathbf{x}_i) = \frac{1}{N} \sum_{j=1}^N \frac{1}{\sqrt{2\pi}h} e^{-\frac{1}{2}\left(\frac{x_{i1} - x_{j1}}{h}\right)^2} \frac{1}{\sqrt{2\pi}h} e^{-\frac{1}{2}\left(\frac{x_{i2} - x_{j2}}{h}\right)^2} \dots \frac{1}{\sqrt{2\pi}h} e^{-\frac{1}{2}\left(\frac{x_{in} - x_{jn}}{h}\right)^2} \quad (21)$$

$$p(\mathbf{x}_i) = \frac{1}{N} \sum_{j=1}^N \frac{1}{(\sqrt{2\pi}h)^n} e^{-\frac{1}{2} \sum_{k=1}^n \left(\frac{x_{ik} - x_{jk}}{h}\right)^2} \quad (22)$$

$$p(\mathbf{x}_i) = \frac{1}{N(\sqrt{2\pi}h)^n} \sum_{j=1}^N e^{-\frac{(\mathbf{x}_i - \mathbf{x}_j)^2}{2h^2}} \quad (23)$$



(a) Amostras de distribuições multimodais e contornos das distribuições aproximadas com o KDE multidimensional.

(b) Função de densidade estimada com o KDE para os dados da Figura 9a

Figura 9: Aproximação de uma função de densidade bivariada com 4 modos utilizando o KDE segundo Equações 19 e 23.

De acordo com a Equação 23 a probabilidade de ocorrência de um vetor arbitrário  $n$ -dimensional  $\mathbf{x}_i$  pode ser estimada, assumindo-se independência entre as  $n$  dimensões, por meio da soma das funções de *kernel* Normais descritas na forma  $g(u_{ij}) = e^{-u_{ij}^2}$  cujos argumentos são as distâncias Euclidianas entre  $\mathbf{x}_i$  e todos os vetores  $\mathbf{x}_j$  do conjunto de amostras, ponderadas por  $h$ . Uma vez obtida a matriz de distâncias  $\mathbf{D} = [d_{ij}]$ , a matriz de kernel gaussiano  $\mathbf{K}$  pode ser diretamente calculada pela aplicação da função  $g(u_{ij})$  a cada elemento  $d_{ij}$  da matriz  $\mathbf{D}$  dividido por  $h$ . O valor da probabilidade de ocorrência de um vetor arbitrário  $\mathbf{x}_i$  pode então ser estimado pela soma da linha (ou coluna)  $i$  da matriz  $\mathbf{K}$ , o que corresponde à Equação 23.

O problema de estimativa da densidade  $p(\mathbf{x}_i)$  segundo a Equação 22 se resume a encontrar então o valor de  $h$  que satisfaça a alguma restrição ou função-objetivo. Para o problema de classificação Bayesiana, por exemplo, a função-objetivo poderia ser a soma dos erros quadráticos do classificador, calculada sobre o conjunto de dados de treinamento. Caso a definição de uma função-objetivo para o problema seja possível, o problema geral do ponto de vista de otimização seria caracterizado como  $\arg \max_h J(\mathbf{K}(\mathbf{x}_i, \mathbf{x}_k), y_i, y_k, h)$ . Como o kernel  $\mathbf{K}(\cdot)$  e os dados são fixos, para o caso da aproximação com o KDE, o único parâmetro a ser ajustado é o raio  $h$ .

## 0.5 Espaço de Verossimilhanças

A aplicação da Regra de Bayes na resolução de um problema de classificação binária, conforme discutido nas seções anteriores, envolve a estimativa das verossimilhanças  $P(\mathbf{x}|C_1)$  e  $P(\mathbf{x}|C_2)$ , cuja razão é comparada com o limiar  $k = \frac{P(C_2)}{P(C_1)}$  para gerar a regra de classificação conforme Equação 14. A equação que determina a separação das duas classes  $C_1$  e  $C_2$  é, portanto, a reta  $P(\mathbf{x}|C_1) - kP(\mathbf{x}|C_2) = 0$  definida no espaço  $P(\mathbf{x}|C_1) \times P(\mathbf{x}|C_2)$ . A regra de classificação da Equação 14 pode então ser re-escrita na forma apresentada na

Equação 24.

$$Classe(\mathbf{x}) = \begin{cases} C_1 & Se\ u \geq 0 \\ C_2 & Caso\ contrário \end{cases} \quad (24)$$

em que  $u = P(\mathbf{x}|C_1) - kP(\mathbf{x}|C_2)$ .

A descrição do classificador binário de Bayes na forma da Equação 24 se assemelha a uma rede neural de duas camadas, conforme representação esquemática da Figura 10.

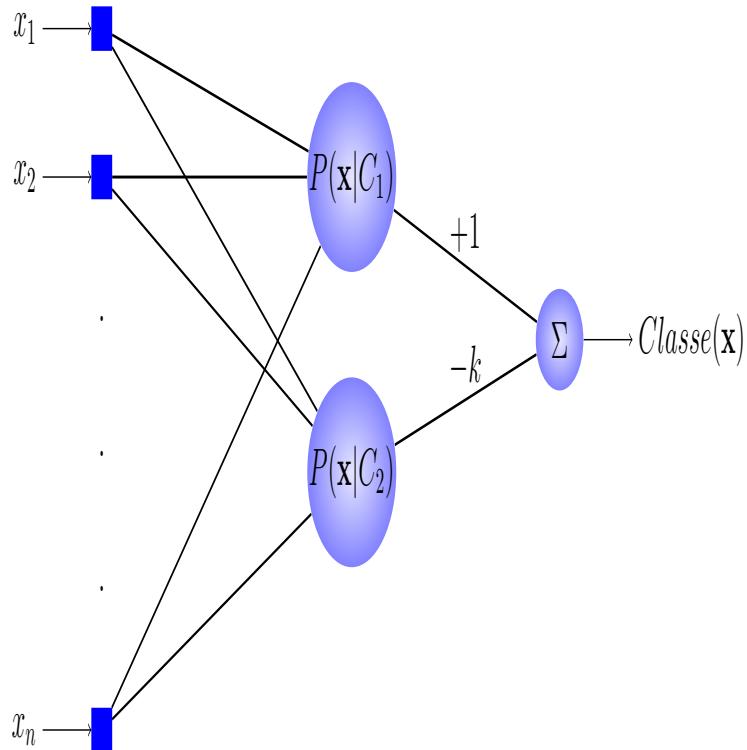


Figura 10: Representação esquemática de um classificador binário de Bayes na forma de uma rede artificial de duas camadas.

Assim, o problema de separação de classes utilizando o classificador Bayesiano envolve uma separação linear envolvendo as verossimilhanças  $P(\mathbf{x}|C_1)$  e  $P(\mathbf{x}|C_2)$ . Este conceito pode ser melhor visualizado por meio do exemplo apresentado na Figura 11. Na Figura 11a, os dados do problema de classificação das espirais são apresentados juntamente com os contornos das funções de densidades das verossimilhanças estimadas com o KDE. Os valores das densidades para cada uma das amostras de treinamento são apresentados na Figura 11b, em que cada ponto  $(P(\mathbf{x}|C_1), P(\mathbf{x}|C_2))$  no gráfico é representado pelo par ordenado contendo os valores das verossimilhanças para cada amostra  $\mathbf{x}$ . Como pode ser observado, o problema de separação não-linear no espaço dos atributos  $\mathbf{x}_1 \times \mathbf{x}_2$ , conforme Figura 11a, tornou-se linear no espaço das verossimilhanças  $P(\mathbf{x}|C_1) \times P(\mathbf{x}|C_2)$ .

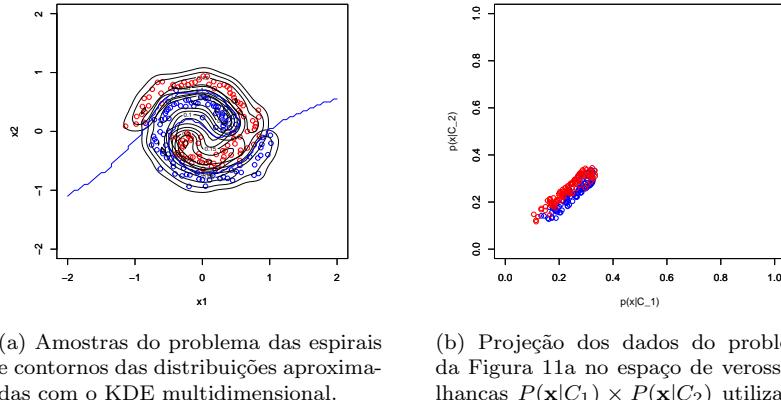


Figura 11: Resolução do problema das espirais utilizando o KDE.

## 0.6 Classificador por matriz de kernel

A visualização da matriz de kernel de um determinado problema bem controlado pode contribuir para a sua análise e entendimento. Considere, portanto, o conjunto de amostras apresentado na Figura 12 e a sua matriz  $\mathbf{K}$  para  $h = 1$  apresentada na Figura 13. A visualização da matriz de kernel nos permite identificar claramente quatro submatrizes distintas que compõem o kernel, as quais serão caracterizadas aqui como  $\mathbf{K}_{11}$ ,  $\mathbf{K}_{12}$ ,  $\mathbf{K}_{21}$  e  $\mathbf{K}_{22}$ . Considerando-se que os dois agrupamentos de dados caracterizam duas classes distintas, as submatrizes  $\mathbf{K}_{11}$  e  $\mathbf{K}_{22}$  contêm as relações intra-classes e as submatrizes  $\mathbf{K}_{12}$  e  $\mathbf{K}_{21}$  contêm as relações entre-classes. Assim a estimativa de densidade de acordo com a Equação 23 pode ser reescrita através da composição das densidades estimadas para matrizes adjacentes, conforme Equações 25 e 26, em que os termos  $P(\{\mathbf{x}_i, y_i = -1\}|C_1)$ ,  $P(\{\mathbf{x}_i, y_i = -1\}|C_2)$ ,  $P(\{\mathbf{x}_i, y_i = +1\}|C_1)$  e  $P(\{\mathbf{x}_i, y_i = +1\}|C_2)$  são descritos a seguir:

- $P(\{\mathbf{x}_i, y_i = -1\}|C_1)$ : Estimativa de  $P(\mathbf{x}_i|C_1)$  para  $y_i = -1$ .
- $P(\{\mathbf{x}_i, y_i = -1\}|C_2)$ : Estimativa de  $P(\mathbf{x}_i|C_2)$  para  $y_i = -1$ .
- $P(\{\mathbf{x}_i, y_i = +1\}|C_1)$ : Estimativa de  $P(\mathbf{x}_i|C_1)$  para  $y_i = +1$ .
- $P(\{\mathbf{x}_i, y_i = +1\}|C_2)$ : Estimativa de  $P(\mathbf{x}_i|C_2)$  para  $y_i = +1$ .

$$P(\mathbf{x}_i \in C_1) = \underbrace{\frac{1}{Nh^m} \sum_{k=1}^{N_1} \mathbf{K}_{11}(\mathbf{x}_i, \mathbf{x}_k)}_{P(\{\mathbf{x}_i, y_i = -1\}|C_1)} + \underbrace{\frac{1}{Nh^m} \sum_{p=1}^{N_2} \mathbf{K}_{12}(\mathbf{x}_i, \mathbf{x}_p)}_{P(\{\mathbf{x}_i, y_i = -1\}|C_2)} \quad (25)$$

$$P(\mathbf{x}_i \in C_2) = \underbrace{\frac{1}{Nh^m} \sum_{k=1}^{N_1} \mathbf{K}_{21}(\mathbf{x}_i, \mathbf{x}_k)}_{P(\{\mathbf{x}_i, y_i = +1\}|C_1)} + \underbrace{\frac{1}{Nh^m} \sum_{p=1}^{N_2} \mathbf{K}_{22}(\mathbf{x}_i, \mathbf{x}_p)}_{P(\{\mathbf{x}_i, y_i = +1\}|C_2)} \quad (26)$$

As verossimilhanças podem então ser estimadas de acordo com as Equações 27, 28, 29 e 30. Em um problema de classificação binária, espera-se que as probabilidades estimadas pelas Equações 27 e 30 sejam maximizadas e aquelas estimadas pelas Equações 28 e 29 sejam minimizadas para cada um dos padrões  $\mathbf{x}_i \in D$ .

$$P(\{\mathbf{x}_i, y_i = -1\}|C_1) = \frac{1}{N_1 h^m} \sum_{k=1}^{N_1} \mathbf{K}_{11}(\mathbf{x}_i, \mathbf{x}_k) \quad (27)$$

$$P(\{\mathbf{x}_i, y_i = -1\}|C_2) = \frac{1}{N_2 h^m} \sum_{k=1}^{N_2} \mathbf{K}_{12}(\mathbf{x}_i, \mathbf{x}_k) \quad (28)$$

$$P(\{\mathbf{x}_i, y_i = +1\}|C_1) = \frac{1}{N_1 h^m} \sum_{k=1}^{N_1} \mathbf{K}_{21}(\mathbf{x}_i, \mathbf{x}_k) \quad (29)$$

$$P(\{\mathbf{x}_i, y_i = +1\}|C_2) = \frac{1}{N_2 h^m} \sum_{k=1}^{N_2} \mathbf{K}_{22}(\mathbf{x}_i, \mathbf{x}_k) \quad (30)$$

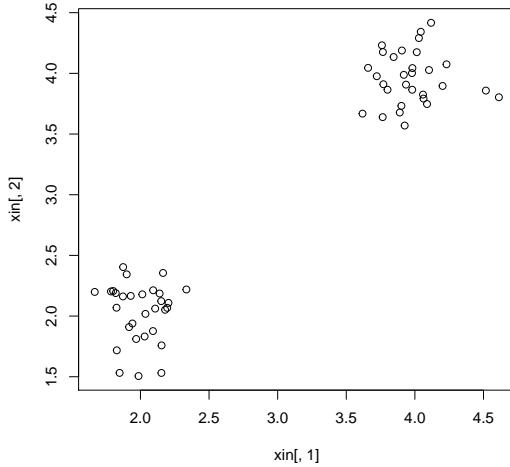


Figura 12: Dados amostrados de duas distribuições Gaussianas com médias em  $m_1 = [2, 2]^T$  e  $m_2 = [4, 4]^T$ .

## 0.7 Regra de Bayes

Considerando-se que os rótulos  $y_i, \forall \mathbf{x}_i \in D$  são conhecidos, espera-se que a estimativa de densidade pelo KDE seja capaz de maximizar as probabilidades posteriores  $P(C_1|\mathbf{x}_i \in C_1)$  e  $P(C_2|\mathbf{x}_i \in C_2)$ . Ao mesmo tempo espera-se que as probabilidades cruzadas  $P(C_1|\mathbf{x}_i \in C_2)$  e  $P(C_2|\mathbf{x}_i \in C_1)$  sejam minimizadas.

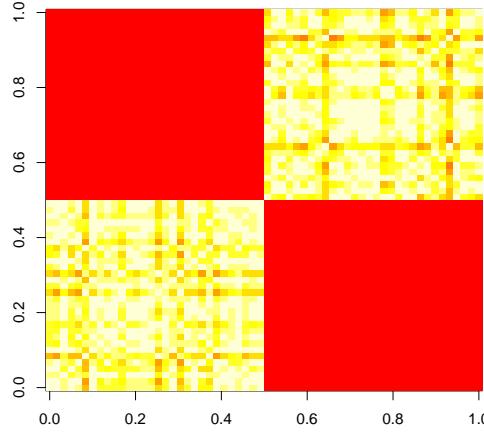


Figura 13: Kernel Gaussiano  $\mathbf{K}$  para o exemplo da Figura 12 com  $h = 1$ .

Assim, de acordo com as Equações 27, 28, 29 e 30 e a classificação pela Regra de Bayes, as desigualdades das Equações 31 e 32 devem ser atendidas.

$$\forall \mathbf{x}_i \in C_1 : N_1 \left( \frac{1}{N_1 h^m} \sum_{k=1}^{N_1} \mathbf{K}_{11}(\mathbf{x}_i, \mathbf{x}_k) \right) - N_2 \left( \frac{1}{N_2 h^m} \sum_{k=1}^{N_2} \mathbf{K}_{12}(\mathbf{x}_i, \mathbf{x}_k) \right) \geq 0 \quad (31)$$

$$\forall \mathbf{x}_i \in C_2 : N_1 \left( \frac{1}{N_1 h^m} \sum_{k=1}^{N_1} \mathbf{K}_{22}(\mathbf{x}_i, \mathbf{x}_k) \right) - N_2 \left( \frac{1}{N_2 h^m} \sum_{k=1}^{N_2} \mathbf{K}_{21}(\mathbf{x}_i, \mathbf{x}_k) \right) \geq 0 \quad (32)$$

O que leva às Equações 33 e 34.

$$\forall \mathbf{x}_i \in C_1 : \frac{1}{h^m} \left( \sum_{k=1}^{N_1} \mathbf{K}_{11}(\mathbf{x}_i, \mathbf{x}_k) - \sum_{k=1}^{N_2} \mathbf{K}_{12}(\mathbf{x}_i, \mathbf{x}_k) \right) \geq 0 \quad (33)$$

$$\forall \mathbf{x}_i \in C_2 : \frac{1}{h^m} \left( \sum_{k=1}^{N_1} \mathbf{K}_{22}(\mathbf{x}_i, \mathbf{x}_k) - \sum_{k=1}^{N_2} \mathbf{K}_{21}(\mathbf{x}_i, \mathbf{x}_k) \right) \geq 0 \quad (34)$$

Mas como  $\frac{1}{h^m}$  será sempre positivo, as restrições das Equações 33 e 34 se reduzem as Equações 35 e 36.

$$\forall \mathbf{x}_i \in C_1 : \sum_{k=1}^{N_1} \mathbf{K}_{11}(\mathbf{x}_i, \mathbf{x}_k) - \sum_{k=1}^{N_2} \mathbf{K}_{12}(\mathbf{x}_i, \mathbf{x}_k) \geq 0 \quad (35)$$

$$\forall \mathbf{x}_i \in C_2 : \sum_{k=1}^{N_1} \mathbf{K}_{22}(\mathbf{x}_i, \mathbf{x}_k) - \sum_{k=1}^{N_2} \mathbf{K}_{21}(\mathbf{x}_i, \mathbf{x}_k) \geq 0 \quad (36)$$

O que nos leva finalmente à Equação 37.

$$\forall \mathbf{x}_i \in D : \sum_{k=1}^N y_i y_k \mathbf{K}(\mathbf{x}_i, \mathbf{x}_k) \geq 0 \quad (37)$$

# Referências Bibliográficas

- [DHS01] R. Duda, P. Hart, and D. Stork. *Pattern Classification*. John Wiley and Sons, 2001. 0-471-05669-3.
- [EM77] Maximum likelihood from incomplete data via the em algorithm. 39(1):1–38, 1977.
- [GBD92] S. Geman, E. Bienenstock, and R. Doursat. Neural networks and the bias-variance dilemma. *Neural Computation*, 4:1–58, 1992.
- [HY01] David J Hand and Keming Yu. Idiot’s bayes—not so stupid after all? *International statistical review*, 69(3):385–398, 2001.
- [Koh95] Ron Kohavi. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence*, pages 1137–1143. Morgan Kaufmann, 1995.
- [KR90] Leonard Kaufman and Peter J. Rousseeuw. *Finding Groups in Data – An Introduction to Cluster Analysis*. John Wiley & Sons, 1990.
- [LG08] A. Leon-Garcia. *Probability, Statistics, and Random Processes for Electrical Engineering*. Pearson/Prentice Hall, 3rd edition, 2008.
- [Mac67] J. MacQueen. Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics*, pages 281–297, Berkeley, Calif., 1967. University of California Press.
- [Par62] Emanuel Parzen. On estimation of a probability density function and mode. *Ann. Math. Statist.*, 33(3):1065–1076, 09 1962.
- [Sil86] B. W. Silverman. *Density estimation: for statistics and data analysis*. London, 1986.